The Ring of Algebraic Functions on Persistence Bar Codes

Aaron Adcock Erik Carlsson Gunnar Carlsson *

April 3, 2013

1 Introduction

Persistent homology ([3], [13]) is a fundamental tool in the area of computational topology. It can be used to infer topological structure in data sets (see [1], [4]), but variations on the method can be applied to study aspects of the shape of point clouds which are not overtly topological ([5], [8]). The methodology assigns to any finite metric space (such as are typically obtained in experimental data of various kinds) and non-negative integer k a *bar code*, by which we will mean a finite collection of intervals with endpoints on the real line. The integer k specifies a dimension of a feature (zero-dimensional for a cluster, one-dimensional for a loop, etc.), and an interval represents a feature which is "born" at the value of a parameter (the persistence parameter) given by the left hand endpoint of the interval, and which "dies" at the value given by the right hand endpoint. These barcodes have been demonstrated to identify structure in spaces of image patches in [1] and [4], and have been demonstrated to distinguish between handdrawn letters in [8]. Because of the unusual structure of the invariant, i.e. as a collection of intervals rather than numerical quantities, the method currently requires substantial knowledge of topological methods. It would clearly be useful to assign and interpret various numerical quantities attached to bar codes, so that these outputs could be used as input to standard algorithms within machine learning, cluster analysis, and other methods. It is the purpose of this

*Research supported in part by NSF DMS-0406992

paper to identify an algebra of functions on the set of bar codes which is defined in a conceptually coherent way.

The main idea is the following. A bar code with exactly n intervals can be specified by a vector $(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$, where x_i denotes the left endpoint of the i-th interval and y_i the right endpoint. However, this representation is many to one, in that the bar code structure does not retain the ordering on the intervals. In fact, the set of bar codes with exactly n intervals can be identified with the set

$$Sp^n(\mathbb{R}^2)$$

the *n-fold symmetric product* of $\mathbb{R}^2$. For any set X, $Sp^n(X)$ is defined to be the orbit space of the action of the symmetric group on n letters on the product X^n given by permuting the coordinates. On the other hand, the space $(\mathbb{R}^2)^n$ is an algebraic variety over $\mathbb{R}$ ([10]). In fact, it is an affine space of dimension $2n$, and the symmetric group action mentioned above is an algebraic action. It is then known (see [12]) that the orbit space inherits the structure of an algebraic variety, and the elements of its *affine coordinate ring* ([10]) are functions on the set of bar codes with exactly n intervals. These affine coordinate rings are well known algebras referred to generically as rings of *multisymmetric polynomials* ([9]). They can be quite complicated, since it turns out that any set of algebra generators for them will satisfy non-trivial relations or *syzygies*. It turns out, though, that there are inclusions of algebraic varieties

$$Sp^n(\mathbb{R}^2) \to Sp^{n+1}(\mathbb{R}^2) \tag{1}$$

which produce an inverse system of graded affine coordinate rings

$$\cdots \to A[Sp^{n+1}(\mathbb{R}^2)] \to A[Sp^n(\mathbb{R}^2)] \to \cdots$$

whose inverse limit we will denote, by abuse of notation, by $A[Sp^\infty(\mathbb{R}^2)]$. The notation $A[-]$ denotes the affine coordinate ring. This algebra is known to be freely generated on a set of minimal algebra generators ([9]).

The analysis of the system (1) above is not sufficient, though. This system identifies a point $((x_1, y_1), \ldots, (x_n, y_n)) \in Sp^n(\mathbb{R}^2)$ with the point

$$((x_1, y_1), \ldots, (x_n, y_n), (0, 0)) \in Sp^{n+1}(\mathbb{R}^2)$$

In other words, a set S of n intervals is identified with the set of $n + 1$ intervals obtained by adjoining the interval of length zero whose two endpoints are zero. However, in the parametrization of the isomorphism classes

of persistence vector spaces in [13] by barcodes, any interval of length zero is identified with the zero module. So, we would like to determine the ring of all algebraic functions (i.e. the elements of $A[Sp^\infty(\mathbb{R}^2)]$) which have the property that they take the same value on any barcode as on the result of adjoining any interval of length zero to it. In this paper, we will identify this subring, describe its structure, and describe the algebra generators explicitly so that they can be used effectively by those interested in analyzing databases of shapes.

2 The Ind-scheme $\mathfrak{B}$

We first discuss the set of bar codes, without any algebraic variety structures. For every n, we first consider the set of bar codes containing exactly n intervals. We will permit intervals of length zero. The set of intervals $\mathfrak{I}$ can be identified with the subset of $\mathbb{R}^2$ consisting of pairs (x, y) with $x \leq y$. A bar code containing n intervals is therefore identified with the n-fold symmetric product $Sp^n(\mathfrak{I})$, where for any set X, $Sp^n(X)$ is defined to be the orbit space of the action of the symmetric group on n letters on the product X^n given by permuting the coordinates. One can assemble these sets into a directed system

$$\mathfrak{I} \stackrel{i_1}{\to} Sp^2(\mathfrak{I}) \stackrel{i_2}{\to} Sp^3(\mathfrak{I}) \stackrel{i_3}{\to} Sp^4(\mathfrak{I}) \to \cdots$$

where the maps $i_n : Sp^n(\mathfrak{I}) \to Sp^{n+1}(\mathfrak{I})$ are given by

$$i_n(\{I_1, \ldots, I_n\}) = \{I_1, \ldots, I_n, [0, 0]\}$$

The direct limit of this system will be denoted by $Sp^\infty(\mathfrak{I})$. We are interested in studying functions on $Sp^\infty(\mathfrak{I})$. Such a function can be identified with an infinite vector $(f_1, f_2, f_3, \ldots)$ of functions $f_n : Sp^n(\mathfrak{I}) \to \mathbb{R}$ satisfying the compatibility condition

$$f_{n+1} \cdot i_n = f_n$$

The set of all such vectors of functions forms a ring $\mathcal{R}$ under coordinate-wise addition and multiplication. It is not exactly what we want, however. The reason is that under the parametrization of persistence vector spaces as described in [3] and [13], intervals of length zero correspond to zero vector spaces, and therefore all intervals of length zero should be considered equal. This means that we should consider only functions $F : Sp^\infty(\mathfrak{I}) \to \mathbb{R}$ for

which

$$F(\{I_1, I_2, \ldots, I_n, [\xi, \xi]\}) = F(\{I_1, I_2, \ldots, I_n, [\eta, \eta]\})$$

for all possible values of ξ and η. The set of all such functions is a subring $\mathcal{R}' \subseteq \mathcal{R}$. This set of functions can be defined as the set of all functions on the set $\mathfrak{B}$ defined by

$$\mathfrak{B} = \coprod_n Sp^n(\mathfrak{I})/\simeq$$

where $\simeq$ is the equivalence relation generated by all relations of the form $\{I_1, I_2, I_n, [\xi, \xi]\} \simeq \{I_1, I_2, \ldots, I_n\}$.

Remark: The reader may suggest that one consider instead only the subset $\mathfrak{I}^+$ consisting of intervals of positive length. This will produce a disjoint union of sets of barcodes, partitioned into the sets containing a fixed positive number of intervals of positive length. Such a description does not take into account the fact that we would like to topologize the space of all bar codes in such a way that

$$lim_{\epsilon \to 0}\{I_1, I_2, \ldots I_n, [x_{n+1}, x_{n+1} + \epsilon]\} = \{I_1, I_2, \ldots, I_n\}$$

The reason for this is that small perturbations to the input data to the persistence algorithms can modify the barcodes by modifying lengths of intervals a small amount and add intervals of small length. This is the stability theorem for persistence diagrams proved in [7].

The ring of functions $\mathcal{R}'$ is too large to deal with effectively. Even the much smaller ring of continuous functions on $\mathfrak{B}$ is still too complex to describe completely. We will observe that $\mathfrak{B}$ is described as a colimit of algebraic varieties, and that it is therefore possible to define the *ring of algebraic functions on* $\mathfrak{B}$. It is this ring we will analyze.

Throughout this paper, k will denote the field $\mathbb{R}$. All varieties will be over k. We consider the affine space $\mathfrak{A}_n = \mathbb{A}(n)$ of dimension $2n$, parametrized with coordinates $(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$. Its affine coordinate ring is the polynomial ring $B_n = k[x_1, y_1, \ldots, x_n, y_n]$. There is an action of the symmetric group S_n on n letters on $\mathfrak{A}_n$, and from [12] it follows that the set of orbits on the set of points of the variety is itself an affine algebraic variety, with affine coordinate ring equal to the invariant subring $B_n^{S_n}$. Let $W_i \subseteq \mathfrak{A}_n$ denote the subvariety $y_i - x_i = 0$. We let $D_n \subseteq B_n$ denote the subring of functions whose restriction to W_i is independent of x_i for all i. We wish to characterize this subring algebraically.

Proposition 1. *The ring D_n is characterized algebraically as the subring of all f for which*

$$(\frac{\partial}{\partial x_i} + \frac{\partial}{\partial y_i})f \in (y_i - x_i)$$

for all i.

Proof: We fix i, and consider all the functions f for which $f|W_i$ is independent of x_i (and therefore y_i). The operator $\frac{\partial}{\partial x_i} + \frac{\partial}{\partial y_i}$ induces a differential operator on the quotient ring $Q_n = B_n/(y_i - x_i)$, which is identified with the partial differential operator $2\frac{\partial}{\partial x_i}$ in

$$Q_n \cong k[x_1, y_1, \ldots, y_{i-1}, x_i, x_{i+1}, \ldots, x_n, y_n]$$

The requirement is that the image $\overline{f}$ of f in W_i is independent of x_i, and this is equivalent to the condition $\frac{\partial}{\partial x_i}(\overline{f}) = 0$. This condition is to hold for each i, which gives the result. □

3 The ring of algebraic functions on $\mathfrak{B}$

We begin by changing coordinates via the formulae $\xi_i = x_i + y_i$ and $\eta_i = y_i - x_i$. It is clear that B_n can also be identified with $k[\xi_1, \eta_1, \ldots, \xi_n, \eta_n]$, and that the symmetric group in the new coordinate system permutes the ξ_i's and η_i's. Under this transformation, the operator $\frac{\partial}{\partial x_i} + \frac{\partial}{\partial y_i}$ is carried into the operator $2\frac{\partial}{\partial \xi_i}$. This means that the ring D_n is identified with the subring of functions $f(\xi_1, \eta_1, \ldots, \xi_n, \eta_n)$ for which $\frac{\partial f}{\partial \xi_i} \in (\eta_i)$ for all i.

Proposition 2. *A k-basis for the D_n is given by the set of monomials*

$$\xi_1^{a_1}\xi_2^{a_2}\cdots\xi_n^{a_n}\eta_1^{b_1}\eta_2^{b_2}\cdots\eta_n^{b_n}$$

for which $a_i > 0$ implies $b_i > 0$.

Proof: We note that the operator $\partial/\partial\xi_i$ carries each monomial to a constant multiple of a single monomial, namely the monomial obtained by decreasing a_i by one. Moreover, containment in the ideal (η_i) is also given purely by conditions on monomials, i.e. that $b_i > 0$. We conclude that D_n is spanned by monomials lying in D_n. But it is clear that a monomial μ lies in D_n exactly if it is the case that whenever ξ_i divides μ, then η_i also divides μ.

This corresponds to the above numerical condition on the exponents in the monomial. □

The symmetric group action clearly preserves the subring D_n. Moreover, it preserves the basis of monomials within D_n. Let $\{\mu_\alpha\}_{\alpha\in A}$ denote a set of orbit representatives of the S_n-action on the set of monomials defined in Proposition 2. Let σ_α denote the sum of all the elements in the orbit of μ_α.

Proposition 3. *We let $D_n^{S_n}$ denote the subring of elements of D_n which are invariant under the action of S_n. Then the elements σ_α form a k-basis of $D_n^{S_n}$.*

Proof: This result plainly holds for any algebra over a field of characteristic zero on which there is a G-action which preserves a basis of monomials. □

We have restriction maps $\pi_{n,m} : D_n \to D_m$, when $n \geq m$, defined by $\pi_{n,m}(\xi_i(\text{resp } \eta_i)) = \xi_i(\text{resp } \eta_i)$ for $i \leq m$, and $\pi_{n,m}(\xi_i) = 0$ for $i > m$. The map $\pi_{n,m}$ is S_m-equivariant, where S_m acts by permuting the first m pairs of variables. It follows that we may construct composites

$$D_n^{S_n} \hookrightarrow D_n^{S_m} \overset{\pi_{n,m}^{S_m}}{\to} D_m^{S_m}$$

which we denote by $\sigma_{n,m}$, and therefore the inverse system

$$\cdots \overset{\sigma_{n+1,n}}{\longrightarrow} D_n^{S_n} \overset{\sigma_{n,n-1}}{\longrightarrow} D_{n-1}^{S_{n-1}} \overset{\sigma_{n-1,n-2}}{\longrightarrow} \cdots \overset{\sigma_{2,1}}{\longrightarrow} D_1$$

We will denote the inverse limit of this graded system by $\mathfrak{D}$.

We next recall some of the notation and basic facts about multisymmetric polynomials, which can be found in Dalbec [9]. Let $R_{n,r}$ be the polynomial ring in nr variables,

$$R_{n,r} = k[x_{11}, x_{21}, ..., x_{nr}],$$

and let

$$\Lambda_{n,r} = R_{n,r}^{S_n},$$

denote the ring of S_n invariants, where the symmetric group acts diagonally. There is an inverse system parallel to the one constructed above involving the rings $\Lambda_{n,r}$. We have evaluation maps

$$\pi_{n,m} : R_{n,r} \to R_{m,r}, \quad m \leq n$$

defined by setting $x_{ir} = 0$ if $i > m$. The map $\pi_{n,m}$ is S_m-equivariant, when $S_m \subseteq S_n$ is the subgroup of permutations of the first m elements of the set $\{1, \ldots, m\}$. We have the composites

$$\Lambda_{n,r} = R_{n,r}^{S_n} \hookrightarrow R_{n,r}^{S_m} \overset{\pi_{n,m}^{S_m}}{\to} R_{m,r}^{S_m} = \Lambda_{m,r}$$

which we denote by $\rho_{n,m}$. The inverse limit of the system

$$\cdots \overset{\rho_{n+1,n}}{\longrightarrow} \Lambda_{n,r} \overset{\rho_{n,n-1}}{\longrightarrow} \Lambda_{n-1,r} \overset{\rho_{n-1,n-2}}{\longrightarrow} \cdots \overset{\rho_{2,1}}{\longrightarrow} \Lambda_{1,r}$$

as graded rings will be denoted by Λ_r, and referred to as the ring of r-*multisymmetric* functions. Its grading is given by

$$\Lambda_r = \bigoplus_k \Lambda_r^k$$

induced by the grading on $R_{n,r}$. There is an evident embedding $\mathfrak{D} \hookrightarrow \Lambda_2$. We will use this embedding to identify the structure of $\mathfrak{D}$.

The ring of multisymmetric functions has several interesting sets of generators. Given some nonzero vectors

$$\mathbf{a}_i = (a_{i1}, ..., a_{ir}) \in \mathbb{N}^r \backslash 0,$$

we define the multisymmetric monomials by

$$m_{\mathbf{a}_1,\ldots,\mathbf{a}_k} = \operatorname{Sym} x_{11}^{a_{11}} \cdots x_{kr}^{a_{kr}} \in \Lambda_r,$$

where Sym is the symmetrization map. Sym applied to a monomial yields the sum of all monomials which are in the orbit of the S_n-action.

They form a vector space basis of $\Lambda_{n,r}$, for any n. It is known that $\Lambda_{n,r}$ is generated as an algebra by the symmetrizations of monomials involving only $\{x_{11}, x_{12}, \ldots, x_{1r}\}$. They are given by the formulae

$$p_{\mathbf{a}} = m_{\mathbf{a}} = \sum_i x_{i1}^{a_1} \cdots x_{ir}^{a_r},$$

and are called the multisymmetric *power sums*. While there are relations among the power sums in finitely many variables, they freely generate the inverse limit Λ_r, making it a polynomial algebra. See [9] for details.

We will be interested in the case $r = 2$. Let us set

$$x_i = x_{i1}, \quad y_i = x_{i2},$$

and let

$$A_n = R_{n,2} = k[x_1, y_1, ..., x_n, y_n].$$

The subalgebra $\mathfrak{D} \subseteq \Lambda_2$ now has the following characterization.

Theorem 1. *As a subalgebra of Λ_2, $\mathfrak{D}$ is freely generated by elements of the form $p_{a+1,b} - p_{a,b+1}$.*

Checking that these generators are contained in B, and that there are no relations between them is easy. The work is in calculating Hilbert series,

$$P(\Omega) = \sum_{k \geq 0} \dim_{\mathbb{R}}(\Omega^k) t^k,$$

with induced grading on Ω. We do this in the following lemmas.

Lemma 4. *An $\mathbb{R}$-basis for B_n is given by the set of monomials*

$$(x_1 - y_1)^{a_1} y_1^{b_1} \cdots (x_l - y_l)^{a_l} y_l^{b_l}$$

for which $b_i > 0$ implies $a_i > 0$.

Proof. Make the substitution $z_i = x_i - y_i$, which corresponds to an isomorphism

$$A = \mathbb{R}[z_1, y_1, ..., z_n, y_n],$$

and notice that B_n is exactly the kernel of the differential operator $\partial/\partial z_i$. The operator $\partial/\partial z_i$ carries each monomial to a constant times a single monomial, namely the monomial obtained by decreasing a_i by one. Moreover, containment in the ideal (z_i) is also given purely by conditions on monomials, i.e. that $a_i > 0$. We conclude that B_n is spanned by monomials lying in B_n. But it is clear that a monomial μ lies in B_n exactly if it is the case that whenever y_i divides μ, then z_i also divides μ. This clearly corresponds to the above numerical condition on the exponents in the monomial. □

Lemma 5. *The Hilbert series of Ω is*

$$P(\Omega) = \prod_{d \geq 1} (1 - t^k)^{-k}.$$

Proof. The above proposition shows that B_n has a basis of monomials which are invariant under the S_n-action. Whenever this is true, any set of orbit representatives constitute a basis of $B_n^{S_n}$ over a field of characteristic zero. We define such a set of representatives by the monomials of the form

$$z_1^{a_1} y_1^{b_1} \cdots z_l^{a_l} y_l^{b_l}, \quad \varphi^{-1}(a_i, b_i) \geq \varphi^{-1}(a_{i+1}, b_{i+1})$$

where $l \leq n$, and $\varphi : \mathbb{N}^+ \to \mathbb{N}^+ \times \mathbb{N}$ is the bijection

$$(\varphi_1, \varphi_2, ...) = ((1,0),(1,1),(2,0),(1,2),(2,1),(3,0),(1,3),...)$$

onto the set of possible nonzero exponents. The dimension of the k-graded component of $B_n^{S_n}$ is just the number of these monomials of degree k.

Let us say that $(a,b) \leq (c,d)$ when $\varphi^{-1}(a,b) \leq \varphi^{-1}(c,d)$, and let $f(a,b,k)$ denote the number of sequences $(a_1, b_1, ..., a_l, b_l)$ such that

$$(a,b) \geq (a_1,b_1) \geq \cdots \geq (a_l,b_l), \quad (a_i,b_i) \in \mathbb{N}^+ \times \mathbb{N}$$

with no restrictions on l. It is easy to check that it satisfies the recursion relation

$$f(a,b,k) = \sum_{(c,d)\leq(a,b)} f(c,d,k-c-d),$$

which leads to the formula

$$\sum_{k\geq 0} f(a,b,k)t^k = (1-t^{a+b})^{-a} \prod_{1\leq k\leq a+b-1} (1-t^k)^{-k}.$$

We then have

$$\lim_{n\to\infty} P(B_n^{S_n}) = \lim_{(a,b)\to\infty} \sum_{k\geq 1} f(a,b,k)t^k = \prod_{k\geq 1}(1-t^k)^{-k}.$$

□

We can now prove the theorem.

Proof. Let

$$\Omega' = \mathbb{R}[p_{10} - p_{01}, p_{11} - p_{20}, ...].$$

It is simple to check that $\Omega' \subset \Omega$. There are no relations between these generators because the homomorphism

$$\Lambda_2 \to \Lambda_2, \quad p_{a+1,b} \mapsto p_{a+1,b} - p_{a,b+1}, \quad p_{0,b} \mapsto p_{0,b}$$

is an isomorphism. It remains to show that the two rings have the same Hilbert series. But $P(\Omega')$ obviously equals the generating function in lemma 5, because there are k generators in degree k. □

4 Machine Learning on $\mathfrak{B}$ with examples

4.1 Digits Example

To illustrate the classification potential of this technique, we apply it to the MNIST database [11], of handwritten digits. We emphasize that the aim is not to outperform existing machine learning algorithms for digit classification, but to present an example that demonstrates one way of combining this technique with existing machine learning techniques. While it is clear that pure topological classification cannot distinguish between the digits (there are three numbers that do not have any loops, three that always have loops, one that has two loops and three that have style-dependent loops), we can use the power of persistent homology to sift out more information. We begin by showing the full analysis of a few digits and then give the empirical results of applying this technique to a subset of the MNIST database.

4.1.1 Topological Methods

We begin by describing a particular graph construction given a digital image. We treat the pixels as vertices and add edges between adjacent pixels (including diagonals). We can now define a filtration on the vertices of the graph corresponding to the image pixels. A natural filtration could be constructed using the pixel intensities of the original image (see Figure 6, Section 4.2). Another filtration, used in [8], can be constructed by thresholding, to produce a binary image, and adding 1-pixels as we sweep across the image. This adds spatial information into what would otherwise be a purely topological measurement. Since the orientation of the digit matters (a 6 is the same as a 9 given a 180 degree rotation), we choose the latter approach and sweep across the rows and columns of each digit.

By taking into account spatial information, we get a rough view of the location of various topological features. For example, though a '9' and '6' both have one connected component and a single loop, the loop will appear at different locations in the top-down filtration for the '9' and '6'. The digits and one of the resulting barcodes are shown in Figures 1 and 2. Using all four sweeps, and both the Betti 0 and Betti 1 barcodes, reveals additional differences between each of the digits.

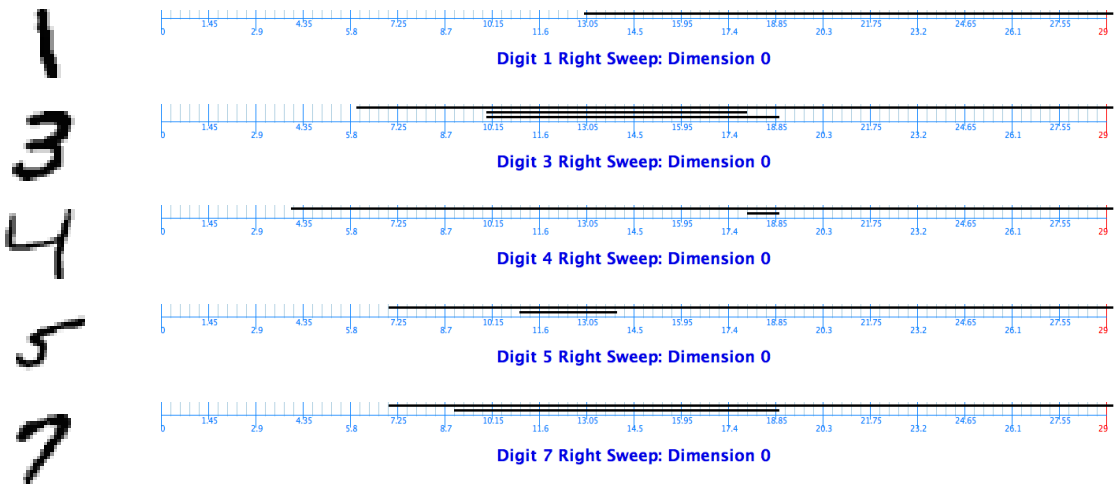


Figure 1: No Loop Digits with Betti 0 barcode, sweep to right

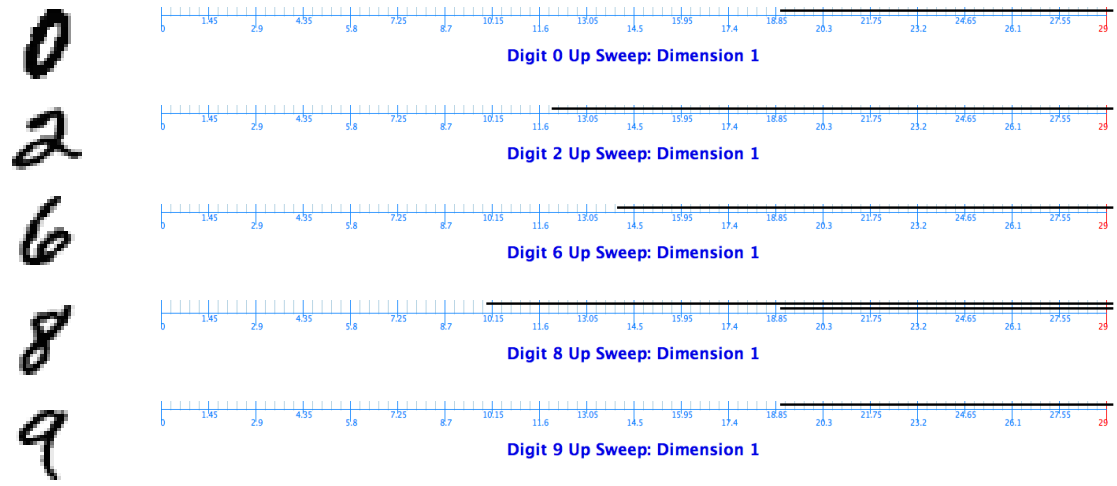


Figure 2: Loop Digits with Betti 1 barcode, sweep to top

4.1.2 Feature Selection

We can use the techniques described in this paper to coordinatize the barcode space $\mathfrak{B}$. In machine learning terminology, these coordinates are called *features*. This allows us to characterize the barcodes generated by each data point as a compact feature vector. This also gives us great flexibility in selecting features that work well with our data. We can then apply a standard

machine learning algorithm, such as a support vector machine (SVM), to classify the data.

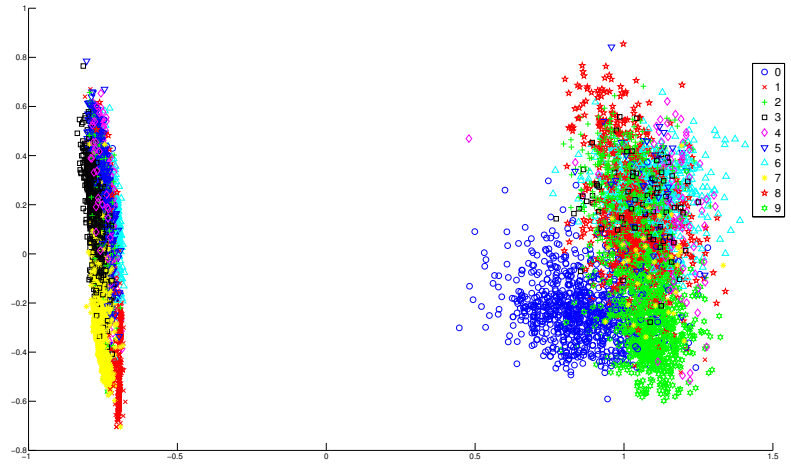
We selected a set of four features from the invariants discussed in this paper. Intuitively, the exponents in each polynomial will give the relative value of small bars or endpoints compared to large bars or endpoints. For example, if comparing two bars of length $\frac{b}{2}$ and b, the first bar will have more weight in an invariant linear polynomial than in an invariant quadratic polynomial. Indeed,

$$\begin{aligned}\left(\frac{b}{2}\right)^2 &= \frac{b^2}{4},\\ \left(\frac{b}{2}\right)^3 &= \frac{b^3}{8},\\ \left(\frac{b}{2}\right)^4 &= \frac{b^4}{16},\\ &\vdots\end{aligned}$$

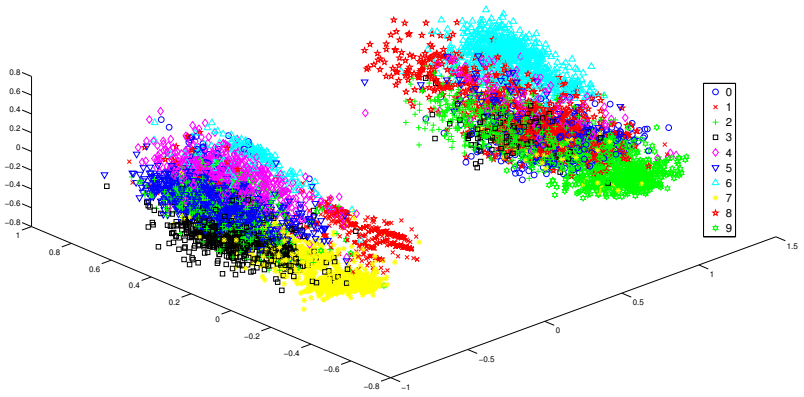
We selected four features,

$$\begin{aligned}&\sum_i x_i(y_i - x_i)\\ &\sum_i (y_{max} - y_i)(y_i - x_i)\\ &\sum_i x_i^2(y_i - x_i)^4\\ &\sum_i (y_{max} - y_i)^2(y_i - x_i)^4\end{aligned}$$

which when applied to the four sweeps, each with a 0-dimensional and 1-dimensional barcode, gives a feature vector of total size 32 which we then arranged into a feature matrix. Intuitively speaking, the first two features take all of the bars, lengths and endpoints, into account. The second two features heavily favor the arrangement of longer bars. A visualization of a matrix of 10,000 digits using classical multidimensional scaling (MDS) is shown in Figure 3 and the spectrum of the matrix is shown in Figure 4.



(a) A 2D View of the Data



(b) A 3D View of the Data

Figure 3: Visualization of Data using Topological Features

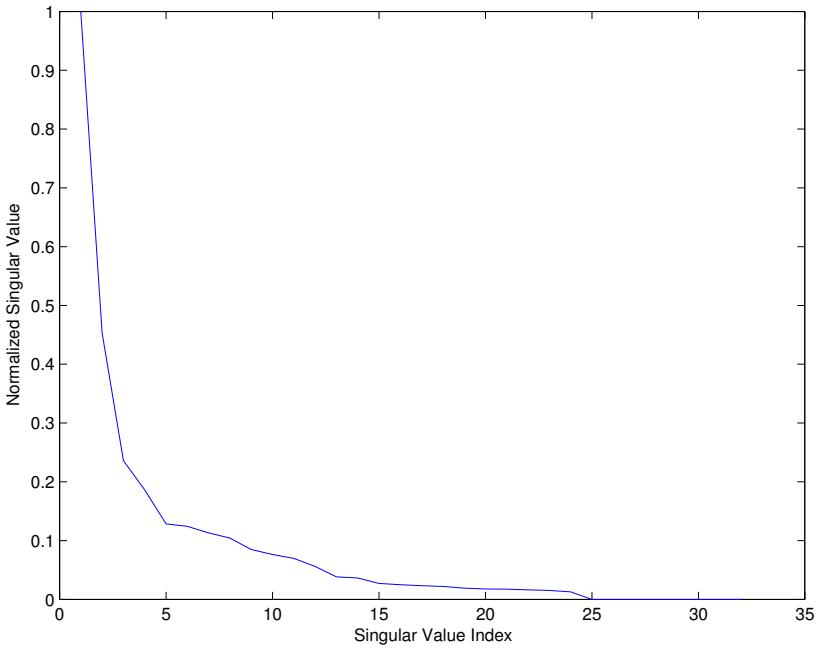


Figure 4: Normalized Spectrum of Topological Feature Matrix

As is typical when using a SVM, we scaled each coordinate such that the values were between 0 and 1. The SVM was implemented using software provided by [6].

4.1.3 Classification Results

We applied these methods on a subset of 1000 digits from the MNIST database to tune parameters of the algorithm and test various kernels. For the radial basis function $e^{-\gamma|u-v|^2}$ (RBF, also known as the Gaussian kernel), we used $\gamma = 8$. For the polynomial kernel $(\gamma(u * v) + a)^d$, we used $d = 3$ with $\gamma = 2$ and $a = 2$. In both functions, u and v represent the calculated feature vectors. After this, we progressively increased the size of the subset to 10,000 handwritten digits.

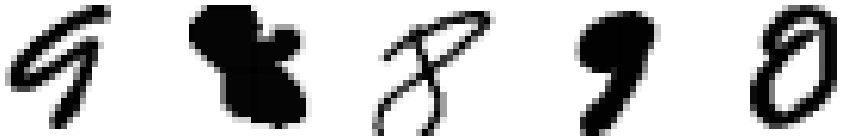
The classification accuracy was measured by partitioning the data set into one hundred subsets and using cross-validation successively on each subset. The results are shown in Table 2.

Table 1: Classification Accuracy of two SVM Kernels

SVM	1000 Digits	5000 Digits	10000 Digits
Gaussian	87.70%	91.54%	92.04%
Polynomial	88.00%	91.62%	92.10%

With the polynomial kernel, an error of 7.9% is seen. As mentioned above, the purpose of this test is not to outperform existing classification algorithms but to demonstrate one application of the topological features. In line with this, we examined some of the digits that the algorithm failed on. Figure 5 shows a few of the typical problem digits.

(a) Stylistic Problems



(b) Spurious Topological Changes

Figure 5: Common Misclassifications

The most common confusion is between a '5' and a '2' written with no loop. Other confusions often occur between the shown style of '7' and slanted '3's and between a certain style of '4' and a '9'. These confusions are not unexpected since these numbers are topologically the same. The extra spatial information added by the directional sweeps is sensitive to variations in the slant or style of handwriting and a visual inspection of these digits suggests why the algorithm has difficulty classifying these particular examples. Other common confusions occur when topological changes occurred to the digit, specifically when the writer adds or removes a loop.

4.2 Hepatic Lesion Classification

In this example, we apply topological features to classifying hepatic lesions. The dataset consists of computed tomography (CT) scans of 132 hepatic lesions that are outlined and annotated by radiologists. There are nine diagnoses represented in the data: cysts (45 lesions), metastases (45 lesions), hemangiomas (18 lesions), hepatocellular carcinomas (HCC, 11 lesions), focal nodules (5 lesions), abscesses (3 lesions), neuroendocrine neoplasms (NeN, 3 lesions), a single laceration and a single fat deposit. Additionally, there are no controls for the size of the lesion and the lesions vary from under 100 pixels to 10,000 pixels. Because of the unbalanced nature of the data, we focus on the subset of cysts, metastases, and hemangiomas.

Classification results using the barcode metric (matching metric) were first presented in [2], and we follow the same methods for processing and generating barcodes from the data. We will briefly describe the methods here. For a more detailed account, please read [2].

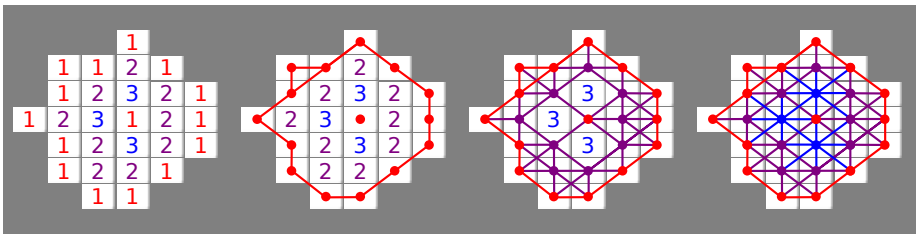
4.2.1 Topological Methods

As mentioned above, a natural filtration for an image is to filter by the pixel intensity. An example of this filtration is given in Figure 6. The variation in pixel intensity allows us to use a one-dimensional filtration on the pixel intensity, but as the results will show, the classification is improved when geometric information is added into the filtrations.

As there is no rotational orientation of the lesions, we cannot add in geometric information using the sweeps described in the previous section. Instead, we use the lesion border provided by the radiologist and assign each pixel its distance from the border. Then, by using two-dimensional homology, we achieve improved results, especially in the case of the hemangiomas which are characterized by large dense regions on the outer part of the lesion. Because two-dimensional filtrations are computationally intensive, we approximate the two-dimensional filtration with one-dimensional barcode 'slices' along the border filtration axis. We use 7 slices per lesion and both the Betti 0 and Betti 1 barcodes.

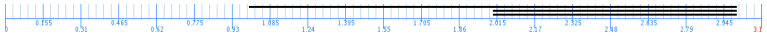
Note that we can look at each filtration from each direction and catch different features. The intensity filtration can add high intensity pixels first or low intensity pixels first. The boundary filtration can begin with pixels near the boundary first or pixels far from the boundary first. This yields 56

one-dimensional barcodes per lesion.



(a) Simple image with filtered complex

(b) β_0 barcode for above image



(c) β_1 barcode for above image

Figure 6: Constructing an increasing 1D-filtration on an image [2]

4.2.2 Feature Selection

We use a slightly different set of four features as compared to the digits example. These features are shown below. The two sets of features that focus on long bars and features which take into account shorter bars is used here. In this application, this is analogous to filtering the barcode to remove the large number of smaller bars. Because of the variations in lesion size, we look at the average over each bar to try and eliminate the effects of large variations in lesion size.

$$\sum_{i}^{n} x_i(y_i - x_i)/n$$
$$\sum_{i}^{n} (y_{max} - y_i)(y_i - x_i)/n$$
$$\sum_{i}^{n} x_i^2(y_i - x_i)^4/n$$
$$\sum_{i}^{n} (y_{max} - y_i)^2(y_i - x_i)^4/n$$

As mentioned above, we have 56 barcodes per lesion. With four features, this yields a feature vector of 224 features for each lesion.

4.2.3 Classification Results

We apply the SVM using only the Gaussian kernel and use an exponential parameter sweep to find optimal values of γ for each method. We use LOOCV to calculate the classification accuracies. The results are shown below. Table 2 gives the results for 1D and 2D filtrations for several different datasets while Table 3 shows how well the algorithm performs on different lesion types for the different filtrations. Table 4 demonstrates the effect of size on classification.

Table 2: SVM Classification Accuracies for 1D and 2D Filtrations

Filtration	Full	HcHeCM	HeCM	CM
1D (Intensity)	53.03%	59.66%	65.74%	75.56%
2D	67.42%	74.79 %	81.48%	86.67%

Using [2], we see that that topological features are comparable with using the matching metric to generate features. The results from the HeCM dataset for the two methods are shown below. They reflect the correct classification of a single lesion using a the topological features, making the two methods

Table 3: HeCM % Classification Accuracy by Lesion Type

Filtration	% of HeCM	% of Heman.	% of Cysts	% of Metas.
1D	65.74%	33.33%	75.56%	68.89%
2D	81.48%	61.11%	86.67%	84.44%

Table 4: Classification by Lesion Size of HeCM

Lesion Size by Area	% Accu.	# of Heman.	# of Cysts	# of Metas.
All	81.48%	18	45	45
<10000 px	82.52%	18	42	43
<5000 px	84.78%	16	39	37
<2500 px	86.25%	14	32	34
<1250 px	88.514%	8	28	23

virtually the same for this subset of the data. Comparing with the other results in [2] shows that the two results are very close in most categories, with each slightly outperforming the other in certain subsets of the data.

Table 5: Classification Methods

Filtration	Barcode Features	Matching Metric
1D	65.74%	63.80%
2D	81.48%	80.56%

4.3 Discussion

These two examples demonstrate the classifying power of topological features when applied to real world datasets. This was done using off-the-shelf machine learning algorithms showing that these features can easily be combined with more traditional classification methods adding a set of additional classification features to the machine learning toolbox.

These examples also show the power of combining topology with geometry. In both datasets, this is an integral part of the classification procedure. The results in the hepatic lesion dataset provide an especially good example of the potential gains that can be achieved by combining both fields.

In summary, using algebraic geometry and invariant theory, we have identified a family of coordinates on the space of finite metric spaces, or sampled shapes. These coordinates can serve as a method for organizing the collection of all barcodes, and therefore any database whose members produce barcodes. Of course, we can also use various metrics on barcode space, such as the bottleneck or Wasserstein distances. It would be extremely interesting to analyze the relationship between these distances on barcode spaces with various more algebraic notions of distance on the barcode coordinates. It would also be very interesting to define and analyze analogous coordinates on spaces of multidimensional persistence modules, where they might give information which is currently not accessible due to the complexity of the algebraic descriptions of multidimensional persistence modules.

References

[1] H. Adams and G. Carlsson, *On the non-linear statistics of range image patches*, SIAM J. Imaging Sci., (2),1, p. 110-117.

[2] A. Adcock, D. Rubin, and G. Carlsson, *Classification of Hepatic Lesions using the Matching Metric*, arXiv preprint arXiv:1210.0866 (2012).

[3] G. Carlsson, *Topology and data*, Bull. Amer. Math. Soc. (46), 2, 2009, pp. 255-308.

[4] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, *On the local behavior of spaces of natural images*, to appear, International Journal of Computer Vision, (76), 1, 2008, pp. 1-12.

[5] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, *Persistence barcodes for shapes*, International Journal of Shape Modeling, (11), (2005), pp. 149-187

[6] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*, ACM Transactions on Intelligent Sys-

tems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[7] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot, *Gromov-Hausdorff stable signatures for shape using persistence*, Computer Graphics Forum, (28), 5, 2009, pp. 1393-1403.

[8] A. Collins, A. Zomorodian, G. Carlsson, and L. Guibas, *A barcode shape descriptor for curve point cloud data*, Computers and Graphics, Volume 28, 2004, pp. 881-894.

[9] J. Dalbec, *Multisymmetric functions*, Beiträge Algebra Geom., (40),1, 1999, pp.27-51.

[10] W. Fulton and R. Weiss, **Algebraic curves: An introduction to algebraic geometry**, Addison-Wesley, 1989.

[11] Y. LeCun and C. Cortes, *The MNIST Database*, Courant Institute NYU, Accessed 16 July 2012, http://yann.lecun.com/exdb/mnist/

[12] D. Mumford, J. Fogarty, and F. Kirwan, **Geometric Invariant Theory**, Springer Verlag, 2002.

[13] A. Zomorodian and G. Carlsson, *Computing persistent homology*, Discrete and Computational Geometry, 33 (2), 2005, pp. 247-274.